

Treecode and fast multipole method for N-body simulation with CUDA

Rio Yokota

Boston University

Lorena A. Barba

Boston University

1 Introduction

The classic N -body problem refers to determining the motion of N particles that interact via a long-distance force, such as gravitation or electrostatics. A straightforward approach to obtaining the forces affecting each particle is the evaluation of all pair-wise interactions, resulting in $\mathcal{O}(N^2)$ computational complexity. This method is only reasonable for moderate-size systems, or to compute near-field interactions, in combination with a far-field approximation. In the previous *GPU Gems* volume [29], the acceleration of the all-pairs computation on GPUs was presented for the case of the gravitational potential of N masses. The natural parallelism available in the all-pairs kernel allowed excellent performance on the GPU architecture, and the direct kernel of [32] achieved over 200 Gigafllops on the GeForce 8800 GTX, calculating more than 19 billion interactions per second with $N=16,384$. In the present contribution, we have addressed the more involved task of implementing the fast N -body algorithms that are used for providing a far-field approximation: the $\mathcal{O}(N \log N)$ treecode [4] and $\mathcal{O}(N)$ fast multipole method [16].

Before embarking on the presentation of the algorithms and how they are efficiently cast onto the GPU, let us give some context. The N -body problem of astrophysics was such a strong motivator to computational science, that it drove creation of a special supercomputer in Japan. The history of this massively successful series of machines, called GRAPE, is summarized in the book by its creators, [28]; a popular science magazine article also gives an overview [35]. The GRAPE machines continued to break records into the 21st century, but the size of the problems they can tackle using the $\mathcal{O}(N^2)$ all-pairs force evaluation is still limited by the computational complexity. As stated in [6], “*complexity trumps hardware.*”

Clever algorithms can have a drastic impact on the capabilities of computational science to solve

challenging problems. A case in point is the fast Fourier transform, which has enabled a variety of successful research areas (*e.g.*, the triumph of spectral methods in the simulation of turbulence). The first viable fast algorithms for N -body problems [2, 4] combined two ideas: (i) approximating the effect of a group of distant particles (charges, or masses) by their first few moments, and (ii) rationally dividing space in a hierarchical fashion to establish acceptable margins of distance for these approximations. These two ideas combined in an algorithm result in the so-called *treecode*, reducing the computational complexity to $\mathcal{O}(N \log N)$. The critical third idea that was introduced in the fast multipole method, FMM, is the “local expansion”. This mathematical representation allows groups of distant particles to interact with *groups* of targets, thereby reducing the complexity further to the ideal $\mathcal{O}(N)$ scaling. One essential difference between treecodes and FMM that remains is the method of achieving a desired accuracy in the approximations. Treecodes ensure a given accuracy by restricting the acceptable distances for group-to-target interactions, while FMM looks to the series representation and chooses a proper truncation for specified accuracy.

The advantage of fast algorithms was appreciated by the GRAPE team early on; a modified treecode by [5] was first used in combination with the GRAPE hardware by [25], and continued in later generations of the machine [27, 21]. The hardware architecture limited the order of the multipole expansions to only the dipole term, however, which motivated the development of a new algorithm: the pseudo-particle method [26]. Thus, the interesting history of the GRAPE project illustrates well the interplay between architecture and algorithms. In fact, there are many parallels with GPUs, as used for general-purpose scientific computing. We are reminded here of the statement in [36]:

“the fundamental law of computer science [is]: the faster the computer, the greater the importance of speed of algorithms.”

Fast algorithms for N -body problems have diverse practical applications. We have mentioned astrophysics, the paradigm problem. Of great importance is also the calculation of electrostatic (Coulomb) interactions of many charged ions in biological molecules. Proteins and their interactions with other molecules constitute a great challenge for computation, and fast algorithms can enable studies at physiologically relevant scales [7, 34]. Both gravitational and electrostatic problems are mathematically equivalent to solving a Poisson equation for the scalar potential. A general method for the solution of Poisson problems in integral form is described in [15], using the FMM in a very interesting way to patch local solutions. In [12], instead, the FMM is applied directly to the volume integral representation of the Poisson problem. These general Poisson solvers based on FMM open the door to using the algorithm in various situations where complex geometries are involved, such as fluid dynamics, and also shape representation and recognition [14].

The FMM for the solution of Helmholtz equations was first developed in [33], and is explained in great detail in the book by [17]. The integral-equation formulation is an essential tool in this context, reducing the volumetric problem into one of an integral over a surface. The FMM allows fast solution of these problems by accelerating the computation of dense matrix-vector products arising from the discretization of the integral problem. In fact, the capability of boundary element methods, BEM,

is in this way significantly enhanced; see [30] and [23]. These developments make possible the use of the FMM for many physical and engineering problems, such as seismic, magnetic and acoustic scattering [?, *e.g.*,]Fujiwara1998,DonepudiETal2003,DarveHave2004b,GumerovDuraismwami2009. The recent book by [24] covers applications in elastostatics, Stokes flow, and acoustics; some notable applications including acoustic fields of building and sound barrier combinations, and also a wind turbine model, were presented in [3].

Due to the variety and importance of applications of treecodes and FMM, the combination of algorithmic acceleration with hardware acceleration can have tremendous impact. Alas, programming these algorithms efficiently is no piece of cake. In this contribution, we aim to present GPU kernels for treecode and FMM in, as much as possible, an uncomplicated, accessible way. The interested reader should consult some of the copious literature on the subject for a deeper understanding of the algorithms themselves. Here, we will offer the briefest of summaries. We will focus our attention on achieving a GPU implementation that is efficient in its utilization of the architecture, but without applying the most advanced techniques known in the field (which would complicate the presentation). These advanced techniques that we deliberately did not discuss in the present contribution are briefly summarized in section 6, for completeness. Our target audience is the researcher involved in computational science with an interest in using fast algorithms for any of the applications mentioned above: astrophysics, molecular dynamics, particle simulation with non-negligible far fields, acoustics, electromagnetics, and boundary integral formulations.

2 Fast N-body simulation

As in [32], we will use as our model problem the calculation of the gravitational potential of N masses. We have the following expressions for the potential and force, respectively, on a body i :

$$\Phi_i = m_i \sum_{j=1}^N \frac{m_j}{r_{ij}}, \quad \mathbf{F}_i = -\nabla \Phi_i \quad (1)$$

Here, m_i and m_j are the masses of bodies i and j , respectively; and $\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i$ is the vector from body i to body j . Since the distance vector \mathbf{r}_{ij} is a function of both i and j , an all-pairs summation must be performed. This results in $\mathcal{O}(N^2)$ computational complexity. In the treecode, the sum for the potential is factored into a near-field and a far-field expansion, in the following way,

$$\Phi_i = \sum_{n=0}^{\infty} \sum_{m=-n}^n m_i r_i^{-n-1} Y_n^m(\theta_i, \phi_i) \sum_{j=1}^N \underbrace{m_j \rho_j^n Y_n^{-m}(\alpha_j, \beta_j)}_{M_n^m}. \quad (2)$$

Calculating the summation for M_n^m in this manner can be interpreted as the clustering of particles in the far field. In the above expression, Y_n^m is the spherical harmonic function, and (r, θ, ϕ) ; (ρ, α, β) are the distance vectors from the center of the expansion to bodies i and j , respectively. The key is to factor the all-pairs interaction into a part that involves only i , and a part that involves only j , hence allowing the summation of j to be performed outside of the loop for i . The condition $\frac{\rho}{r} < 1$,

which is required for the series expansion to converge, prohibits the clustering of particles in the near field. Therefore, a tree structure is used to form a hierarchical list of $\log N$ cells that interact with N particles. This results in $\mathcal{O}(N \log N)$ computational complexity.

The complexity can be further reduced by considering cluster-to-cluster interactions¹. In the FMM, a second series expansion is used for such interactions:

$$\Phi_i = \sum_{n=0}^{\infty} \sum_{m=-n}^n m_i r_i^n Y_n^m(\theta_i, \phi_i) \sum_{j=1}^N \underbrace{m_j \rho_j^{-n-1} Y_n^{-m}(\alpha_j, \beta_j)}_{L_n^m}, \quad (3)$$

where the near-field expansion and far-field expansion are reversed. The condition for this expansion to converge is $\frac{r}{\rho} < 1$, which means that the clustering of particles using L_n^m is only valid in the near field. The key here is to translate multipole expansion coefficients M_n^m of cells in the far field to local expansion coefficients L_n^m of cells in the near field, resulting in a cell-cell interaction. Due to the hierarchical nature of the tree structure, each cell needs to only consider the interaction with a constant number of neighboring cells. Since the number of cells is of $\mathcal{O}(N)$, the FMM has a complexity of $\mathcal{O}(N)$. Also, it is easy to see that keeping the number of cells proportional to N results in an asymptotically constant number of particles per cell. This prevents the direct calculation of the near field from adversely affecting the asymptotic behavior of the algorithm.

The flow of the treecode/FMM calculation is illustrated in Figure 1. This schematic shows how the information of all source particles is propagated to a particular set of target particles. The purpose of this figure is to introduce the naming conventions we use for the 7 distinct operations (P2P, P2M, M2M, M2P, M2L, L2L, L2P, P2P), and to associate these steps to a graphical representation. These naming conventions and graphical representations are used later to describe the GPU implementation and to assess its performance. The difference between the treecode and FMM can be explained concisely using this illustration.

First, the mass/charges of the particles are aggregated into the multipole expansions by calculating M_n^m at the center of all cells (the P2M operation). Next, the multipole expansions are further clustered by translating the center of each expansion to a larger cell and adding their contributions at that level (M2M operation). Once the multipole expansions at all levels of the tree are obtained, the treecode calculates Eq. (2) to influence the target particles directly (the M2P operation). In contrast, the FMM first transforms the multipole expansions to local expansions (M2L operation), and then translates the center of each expansion to smaller cells (L2L operation). Finally, the influence of the far field is transmitted from the local expansions to the target particles by calculating Eq. (3) in the L2P operation. The influence of the near field is calculated by an all-pairs interaction of neighboring particles (P2P). In the present contribution, all of the above operations are implemented as GPU kernels.

The schematic in Fig. 1 shows 2D representations of the actual 3D domain sub-divisions. There are

¹ The groups or clusters of bodies reside in a sub-division of space for which various authors use the term “box” or “cell”; *e.g.*, “leaf-cell” as used in [32] corresponds to the smallest sub-domain.

information moves from red to blue

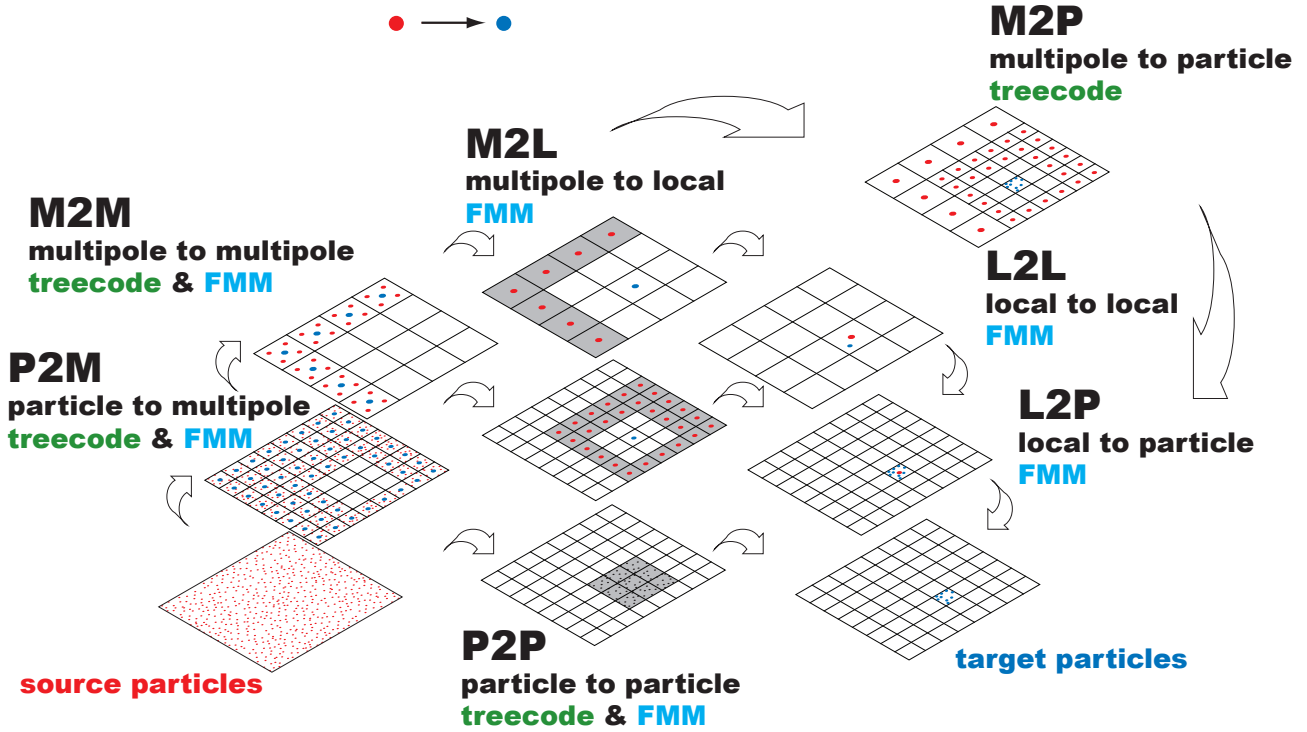


Figure 1: Flow of the treecode and FMM calculation.

two levels of cell division shown, one with 16 cells and another with 64 cells. For a typical calculation with millions of particles, the tree is further divided into 5 or 6 levels (or more). Recall that the number of cells must be kept proportional to the number of particles for these algorithms to achieve their asymptotic complexity. When there are many levels in the tree, the M2M and L2L operations are performed multiple times to propagate the information up and down the tree. Also, the M2L and M2P operations are calculated at every level. The P2M, L2P, and P2P are only calculated at the finest (leaf) level of the tree. Since the calculation load decreases exponentially as we move up the tree, the calculation at the leaf level dominates the work load. In particular, it is the M2L/M2P and P2P that consume most of the runtime in an actual program.

3 CUDA Implementation of the Fast N-body Algorithms

In our GPU implementation of the treecode and FMM algorithms we aim for consistency with the N -body example of [32]. Thus, we will utilize their concept of a computational *tile*: a grid consisting of p rows and p columns representing a subset of the pair-wise interactions to be computed. Consider Fig. 2, which is adapted from a similar diagram used by the previous authors. Each subset of target particles will be handled by different thread blocks in parallel; the parallel work corresponds to the rows on the diagram. Each subset of source particles is sequentially handled by all thread blocks in

chunks of p , where p is the number of threads per thread block. As explained in [32]: “Tiles are sized to balance parallelism with data reuse. The degree of parallelism (that is, the number of rows) must be sufficiently large so that multiple warps can be interleaved to hide latencies in the evaluation of interactions. The amount of data reuse grows with the number of columns, and this parameter also governs the size of the transfer of bodies from device memory into shared memory. Finally, the size of the tile also determines the register space and shared memory required.”

The particle-to-particle (P2P) interactions of the treecode and FMM are calculated in a similar manner (see Figure 3). The entire domain is decomposed into an oct-tree, and each cell at the leaf-level is assigned to a thread block. When the number of particles per cell is larger than the size of the thread block, it is split into multiple thread blocks. The main difference with an all-pairs interaction is that each thread block has a different list of source particles. Thus, it is necessary for each thread block to have its unique index list for the offset of source particles. Only the initial offset (for the cells shown in purple in Figure 3) is passed to the GPU, and the remaining offsets are determined by increments of p .

In order to ensure coalesced memory access, we accumulate all the source data into a large buffer. On the CPU, we perform a loop over all interaction lists as if we were performing the actual kernel execution, but instead of calculating the kernel we store the position vector and mass/charge into one large buffer that is passed on to the GPU. This way, the memory access within the GPU kernel is always contiguous, because the variables are being stored in exactly the same order that they will be accessed. The time it takes to copy the data into the buffer is less than 1% of the entire calculation. Subsequently, the GPU kernel is called and all the information in the buffer is processed in one call (if it fits in the global memory of the GPU). The buffer is split up into an optimum size if it becomes too large to fit on the global memory. We also create a buffer for the target particles, which contains the position vectors. Once they are passed to the GPU, the target buffer will be accessed in strides of p , assigning one particle to each thread. Since the source particle list is different for each target cell (see Figure 3), having particles from two different cells in one thread block causes branching of the instruction. We avoid this by padding the target buffer, instead of accumulating the particles in the next cell. For example, if there are 2000 particles per box and the thread block size is 128, the target buffer will be padded with 48 particles so that it uses 16 thread blocks of size 128 ($16 \cdot 128 = 2048$) for that cell. In such a case, 1 out of the 16 thread blocks will be doing 37.5% excess work, which is an acceptable trade-off to avoid branching of the instruction within a thread block.

The implementation model used for the P2P calculation can be applied to all other steps in the FMM. An example for the M2L translation kernel is shown in Figure 4. Instead of having particle information in each cell, the cell-cell interactions contain many expansion coefficients per cell. Thus, it is natural to assign one target expansion coefficient to each thread while assigning the cell itself to a thread block. Since the typical number of expansion coefficients is in the order of 10-100, the padding issue discussed in the previous paragraph has greater consequences for this case. In the simplest CUDA implementation that we wish to present in this contribution, we simply reduce the thread block size p to alleviate the problem. In the case of particle-cell interactions (P2M) or cell-particles interactions (M2P, L2P), the same logic is applied where either the target expansion coefficients or target particles

are assigned to each thread, and the source expansion coefficients or source particles are read from the source buffer in a coalesced manner and sequentially processed in strides of p .

4 Improvements of Performance

We consider the performance of the treecode and FMM on GPUs for the same model problem as in [32]. We would like to point out that the performance metrics shown here apply for the very basic and simplified versions of these kernels. The purpose of this contribution is to show the reader how easy it is to write CUDA programs for the treecode and FMM. Therefore, many advanced techniques, which would be considered standard for the expert in these algorithms, are deliberately omitted (see section 6). The performance is reported to allow the reader to reproduce the results and verify that their code is performing as expected, and to motivate the discussion about the importance of fast algorithms; we do not claim that the kernels here are as fast as they could be. The CPU tests were run on an Intel Core i7 2.67 GHz, and the GPU tests on an NVIDIA 295GTX. The gcc-4.3 compiler with option `-O3` was used to compile the CPU codes and `nvcc` with `-use_fast_math` was used to compile the CUDA codes.

Figure 5 shows the calculation time against the number of bodies for the direct evaluation, treecode and FMM on a CPU and GPU. The direct calculation is about 300 times faster on the GPU, compared to the single-core CPU. The treecode and FMM are approximately 100 and 30 times faster on the GPU, respectively. For $N < 10^4$, the overhead in the tree construction degrades the performance of the GPU versions. The crossover point between the treecode and direct evaluation is 3×10^3 on the CPU and 2×10^4 on the GPU; the crossover point between the FMM and direct evaluation is 3×10^3 on the CPU and 4×10^4 on the GPU. Note that both for the treecode and FMM, the number of particles at the leaf-level of the tree is higher on the GPU, to obtain a well-balanced calculation (*i.e.*, comparable time should be spent on the near field and on the far field). The crossover point between the treecode and FMM is 3×10^3 on the CPU, but is unclear on the GPU, for the range of our tests.

When the treecode and FMM are performed on the CPU, the P2P and M2P/M2L consume more than 99% of the execution time. When these computationally-intensive parts are executed on the GPU, the execution times of the other stages are no longer negligible. This can be seen in the breakdown shown in Figure 6 for the $N = 10^7$ case. The contribution of each stage is stacked on top of one another, so the total height of the bar is the total execution time. The legend on the left and right correspond to the treecode and FMM, respectively; “sort” indicates the time it takes to reorder the particles so that they are contiguous within each cell; “other” is the total of everything else, including memory allocation, tree construction, interaction list generation, *etc.* The “sort” and “other” operations are performed on the CPU. The depth of the tree in this benchmark is the same for both the treecode and FMM.

As shown in Figure 6, the P2P takes the same amount of time for the treecode and FMM. This is due to the fact that we use the same neighbor list for the treecode and FMM. It may be worth noting that the standard treecode uses the distance between particles to determine the clustering threshold (for a given desired accuracy), and has an interaction list that is slightly more flexible than that of

the FMM. A common measure to determine the clustering in treecodes is the Barnes-Hut multiple acceptance criteria (MAC) $\theta > l/d$ [4], where l is the size of the cell, and d is the distance between the particle and center of mass of the cell. The present calculation uses the standard FMM neighbor list shown in Figure 1 for both the FMM and treecode, which results in a MAC of $\theta = 2/3$. The P2M operation takes longer for the FMM because the order of multipole expansions is larger than in the treecode, to achieve the same accuracy. The calculation loads of M2M, L2L and L2P are small compared to the M2P and M2L. The M2P has a much larger calculation load than the M2L, but it has more data-parallelism. Therefore, the GPU implementation of these two kernels has a somewhat similar execution time. The high data-parallelism of the M2P is an important factor we must consider when comparing the treecode and FMM on GPUs.

Figure 7 shows the measured performance on the GPU measured in Gflop/s; this is actual operations performed in the code, *i.e.*, a `sqrt` counts 1, etc. Clearly, for $N = 10^4$ the GPU is underutilized, but performance is quite good for the larger values of N . The P2P operation performs very well, achieving in the order of 300 Gflop/s for the larger values of N of these tests. The M2P performs much better than the M2L, due to the higher inherent parallelism. This explains why we see the treecode accelerating better overall, compared to FMM, on Figure 5.

5 Detailed description of the GPU kernels

In this section, we give a detailed explanation of the implementation of the treecode/FMM in CUDA. The code snippets shown here are extracted directly from the code available from the distribution released with this article². In particular, we will describe the implementation of the P2P and M2L kernels, which take up most of the calculation time.

5.1 The P2P kernel implementation

We start with the simplest kernel for the interaction of a single pair of particles, shown in Listing 1. Equation (1) is calculated here without the m_i . In other words, it is the acceleration $a_i = F_i/m_i$ that is being calculated. This part of the code is very similar to that of the `nbody` example in the CUDA SDK, which is explained in detail in [32]. The only difference is that the present kernel uses the reciprocal square-root function instead of a square-root and division. There are 19 floating-point operations in this kernel, counting the 3 additions, 6 subtractions, 9 multiplications, and 1 reciprocal square-root. The list of variables is as follows:

- `posTarget` is the position vector of the target particles; it has a `float3` data type and is stored in registers.
- `sharedPosSource` is the position vector and the mass of the source particles; it has a `float4` data type and resides in shared memory.
- `accel` is the acceleration vector of the target particles; it has a `float3` data type and is stored in registers.
- the `float3` data type is used to store the distance vectors `dist`.

²All source code can be found in <http://code.google.com/p/gemsfmm/>

Listing 1: P2P kernel for a single interaction

```

1  __device__ float3 p2p_kernel_core(float3 accel,
2                                     float3 posTarget, float4 sharedPosSource)
3  {
4      float3 dist;
5      dist.x = posTarget.x - sharedPosSource.x;
6      dist.y = posTarget.y - sharedPosSource.y;
7      dist.z = posTarget.z - sharedPosSource.z;
8      float invDist = rsqrtf(dist.x * dist.x + dist.y * dist.y + dist.z * dist.z + eps);
9      float invDistCube = invDist * invDist * invDist;
10     float s = sharedPosSource.w * invDistCube;
11     accel.x -= dist.x * s;
12     accel.y -= dist.y * s;
13     accel.z -= dist.z * s;
14     return accel;
15 }

```

- `eps` is the softening factor [?, see]Aarseth2003.

The function shown in Listing 1 is called from an outer kernel which calculates the pairwise interactions of all particles in the P2P interaction list. This outer kernel is shown in Listing 2, and its graphical representation is shown in Figure 3. The input variables are `deviceOffset`, `devicePosTarget`, `devicePosSource`, and the output is `deviceAccel`. The description of these variables is as follows:

- `deviceOffset` contains the number of interacting cells and the offset of the particle index for each of these cells;
- `devicePosTarget` contains the position vector of the target particles;
- `devicePosSource` is the position vector of the source particles, and
- `deviceAccel` is the acceleration vector of target particles.

All variables that begin with “device” are stored in the device memory. All variables that begin with “shared” are stored in shared memory. Everything else is stored in the registers. Lines 4–10 are declaration of variables; it is possible to reduce register space usage by reusing some of these variables, but for pedagogical purposes we have chosen to declare each variable that has a different functionality. There are 4 variables that are defined externally. One is the `threadsPerBlockTypeA`, which is the number of threads per thread-block for the P2P kernel. We use a different number of threads per thread-block, `threadsPerBlockTypeB`, for the other kernels that have expansion coefficients as targets. On line 5, `threadsPerBlockTypeA` is passed to `threadsPerBlock` as a constant. Another external variable is used on line 7, where `maxP2PInteraction` (the maximum number of neighbor cells in a P2P interaction) is used to calculate `offsetStride` (the stride of the data in `deviceOffset`). The other two externally defined variables are `threadIdx` and `blockIdx`, which are thread index and thread-block index provided by CUDA.

Listing 2: The entire P2P kernel

```

1  __global__ void p2p_kernel(int* deviceOffset, float3* devicePosTarget,
2                               float4* devicePosSource, float3* deviceAccel)
3  {
4      int jbase, jsize, jblok, numInteraction;
5      int j, ij, jj, jb;
6      const int threadsPerBlock = threadsPerBlockTypeA;
7      const int offsetStride = 2 * maxP2PInteraction + 1;
8      float3 posTarget;
9      float3 accel = {0.0f, 0.0f, 0.0f};
10     __shared__ float4 sharedPosSource[threadsPerBlock];
11     posTarget = devicePosTarget[blockIdx.x * threadsPerBlock + threadIdx.x];
12     numInteraction = deviceOffset[blockIdx.x * offsetStride];
13     for(ij = 0; ij < numInteraction; ij++){
14         jbase = deviceOffset[blockIdx.x * offsetStride + 2 * ij + 1];
15         jsize = deviceOffset[blockIdx.x * offsetStride + 2 * ij + 2];
16         jblok = (jsize + threadsPerBlock - 1) / threadsPerBlock;
17         for(j = 0; j < jblok-1; j++){
18             jb = jbase + j * threadsPerBlock + threadIdx.x;
19             sharedPosSource[threadIdx.x] = devicePosSource[jb];
20             __syncthreads();
21         #pragma unroll 32
22             for(jj = 0; jj < threadsPerBlock; jj++){
23                 accel = p2p_kernel_core(accel, posTarget, sharedPosSource[jj]);
24             }
25             __syncthreads();
26         }
27         jb = jbase + j * threadsPerBlock + threadIdx.x;
28         sharedPosSource[threadIdx.x] = devicePosSource[jb];
29         __syncthreads();
30         for(jj = 0; jj < jsize - (j * threadsPerBlock); jj++){
31             accel = p2p_kernel_core(accel, posTarget, sharedPosSource[jj]);
32         }
33         __syncthreads();
34     }
35     deviceAccel[blockIdx.x * threadsPerBlock + threadIdx.x] = accel;
36 }

```

On line 11, the position vectors are copied from the global memory to the registers. On line 12, the number of interacting cells is read from the `deviceOffset`, and on line 13 this number is used to form a loop that goes through all the interacting cells (27 cells for the P2P interaction). Note that each thread block handles (part of) only one target cell, and the interaction list of the neighboring cells is identical for all threads within the thread block. In other words, `blockIdx.x` identifies which target cell we are looking at, and `ij` identifies which source cell it is interacting with. On line 14, the offset of the particle index for that source cell is copied from `deviceOffset` to `jbase`. On line 15, the number of particles in the source cell is copied to `jsize`. Now we have the information of the target particles

and the offset and size of the source particles that they interact with. At this point, the information of the source particles still resides in the device memory. This information is copied to the shared memory in coalesced chunks of size `threadsPerBlock`. However, the number of particles per cell is not always a multiple of `threadsPerBlock`, so the last chunk will contain a remainder that is different from `threadsPerBlock`. It is inefficient to have a conditional branching to detect if the chunk is the last one or not, and it is a waste of storage to pad for each source cell. Therefore, on line 16 the number of chunks `jblock` is calculated by rounding up `jsize` to the nearest multiple of `threadsPerBlock`. On line 17, a loop is executed for all chunks except the last one. The last chunk is processed separately on lines 27–33. On line 18, the index of the source particle on the device memory is calculated by offsetting the thread index first by the chunk offset `j*threadsPerBlock` and then by the cell offset `jbase`. On line 19, this global index is used to copy the position vector of the source particles from device memory to shared memory. Subsequently, `__syncthreads()` is called to ensure that the copy to shared memory has completed on all threads before proceeding. On lines 21–24, a loop is performed for all elements in the current chunk of source particles, where the `p2p_kernel_core` is called per pairwise interaction. The `#pragma unroll 32` is the same loop unrolling suggested in [32]. On line 25, `__syncthreads()` is called to keep `sharedPosSource` from being overwritten for the next chunk before having been used in the current one. Lines 27–33 are identical to lines 18–25 except for the loop counter for `jj`, which is the remainder instead of `threadsPerBlock`. On line 35, the acceleration vector in registers is copied back to the device memory by offsetting the thread index by `blockIdx.x * threadsPerBlock`.

5.2 The M2L kernel implementation

As shown in Equations (2) and (3), the multipole-to-local translation in the FMM is the translation of the multipole expansion coefficients M_n^m in one location to the local expansion coefficients L_n^m at another. If we relabel the indices of the local expansion matrix to L_j^k , the M2L translation can be written as

$$L_j^k = \sum_{n=0}^{p-1} \sum_{m=-n}^n \frac{M_n^m i^{|k-m|-|k|-|m|} A_n^m A_j^k Y_{j+n}^{m-k}(\alpha, \beta)}{(-1)^j A_{j+n}^{m-k} \rho^{j+n+1}} \quad (4)$$

where i is the imaginary unit, p is the order of the series expansion, A_n^m is defined as

$$A_n^m = \frac{1}{\sqrt{(n-m)!(n+m)!}} \quad (5)$$

and Y_n^m is the spherical harmonic

$$Y_n^m(\alpha, \beta) = \sqrt{\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos \alpha) e^{im\beta}. \quad (6)$$

In order to calculate the spherical harmonics, the value of the associated Legendre polynomials P_n^m must be determined. The associated Legendre polynomials have a recurrence relation, which require only the information of $x = \cos \alpha$ to start. The recurrence relations and identities used to generate

Listing 3: Calculation of the spherical harmonic for the M2L kernel

```

1  __device__ void m2l_calculate_ynm(float* sharedYnm,
2                                     float rho, float alpha, float* sharedFactorial)
3  {
4      int i, m, n;
5      float x, s, fact, pn, p, p1, p2, rhom, rhon;
6      x = cosf(alpha);
7      s = sqrt(1 - x * x);
8      fact = 1;
9      pn = 1;
10     rhom = 1.0 / rho;
11     for(m = 0; m < 2 * numExpansions; m++){
12         p = pn;
13         i = m * (m + 1) / 2 + m;
14         sharedYnm[i] = rhom * p;
15         p1 = p;
16         p = x * (2 * m + 1) * p;
17         rhom /= rho;
18         rhon = rhom;
19         for(n = m + 1; n < 2 * numExpansions; n++){
20             i = n * (n + 1) / 2 + m;
21             sharedYnm[i] = rhon * p * sharedFactorial[n - m];
22             p2 = p1;
23             p1 = p;
24             p = (x * (2 * n + 1) * p1 - (n + m) * p2) / (n - m + 1);
25             rhon /= rho;
26         }
27         pn = -pn * fact * s;
28         fact = fact + 2;
29     }
30 }

```

the full associated Legendre polynomial are,

$$(n - m + 1)P_{n+1}^m(x) = x(2n + 1)P_n^m(x) - (n + m)P_{n-1}^m(x), \quad (7)$$

$$P_m^m(x) = (-1)^m(2m - 1)!(1 - x^2)^{m/2}, \quad (8)$$

$$P_{m+1}^m = x(2m + 1)P_m^m(x) \quad (9)$$

The M2L kernel calculates Equation (4) in two stages. First, $Y_n^m/\rho^{n+1}/A_n^m$ is calculated using Equations (5)–(9). Then, Equation (4) is calculated by substituting this result after switching the indices $n \rightarrow j + n$ and $m \rightarrow m - k$. Thus, $M_n^m i^{|k-m|-|k|-|m|} A_n^m A_j^k / (-1)^j$ is calculated at the second stage. Furthermore, in the GPU implementation the complex part $e^{im\beta}$ in Equation (6) is multiplied at the end of the second stage so that the values remain real until then. At the end of the second stage, we simply put the real and complex part of the L_j^k into two separate variables.

The GPU implementation of the first part for $Y_n^m/\rho^{n+1}/A_n^m$ is shown in Listing 3. As was the

case with Listing 1, this function is called from an outer function that calculates the entire M2L translation for all cells. The inputs are `rho`, `alpha`, and `sharedFactorial`. The output is `sharedYnm`. Since, we do not calculate the $e^{im\beta}$ part of the spherical harmonic at this point, `beta` is not necessary. `sharedFactorial` contains the values of the factorials for a given index, *i.e.* `sharedFactorial[n] = n!`. Also, it is $Y_n^m/\rho^{n+1}/A_n^m$ that is stored in `sharedYnm` and not Y_n^m itself. Basically, Equation (7) is calculated on line 24, Equation (8) is calculated on line 27, and Equation (9) is calculated on line 16. `p`, `p1`, and `p2` correspond to P_{n+1}^m , P_n^m , and P_{n-1}^m , respectively. However, `p` is used in lines 14 and 21 before it is updated on lines 16 and 24, so it represents P_n^m at the time of usage. This P_n^m is used to calculate $Y_n^m/\rho^{n+1}/A_n^m$ on lines 14 and 21, although the correspondence to the equation is not obvious at first hand. The connection to the equation will become clear when we do the following transformation,

$$\frac{Y_n^m}{\rho^{n+1}A_n^m} = \frac{\sqrt{(n-m)!/(n+m)!}P_n^m e^{im\beta}}{\rho^{n+1}/\sqrt{(n-m)!(n+m)!}} = \frac{(n-m)!P_n^m}{\rho^{n+1}}e^{im\beta} \quad (10)$$

As mentioned earlier, we do not calculate the $e^{im\beta}$ at this point so `sharedYnm` is symmetric with respect to the sign of m . Therefore, the present loop for the recurrence relation is performed for only $m \geq 0$ and the absolute sign for m in Equation (6) disappears. We can also save shared memory consumption by storing only the $m \geq 0$ half of the spherical harmonic in `sharedYnm`.

The second stage of the M2L kernel is shown in Listing 4. The inputs are `j`, `beta`, `sharedFactorial`, `sharedYnm`, and `sharedMnmSource`. The output is `LnmTarget`. In this second stage of the M2L, the remaining parts of Equation (4) are calculated to obtain L_j^k . Each thread handles a different coefficient in L_j^k . In order to do this, we must associate the `threadIdx.x` to a pair of `j` and `k`. In the outer function, which will be shown later, the index `j` corresponding to `threadIdx.x` is calculated and passed to the present function. Lines 9–11, determine the index `k` from the input `j` and `threadIdx.x`.

We will remind the reader again that this part of the M2L kernel calculates $M_n^m i^{|k-m|-|k|-|m|} A_n^m A_j^k / (-1)^j$. This results in a quadruple loop over the indices j , k , m , and n . However, in the GPU implementation the first two indices are thread-parallelized, only leaving m and n as sequential loops starting from lines 13, 14, and 28. Lines 14–27 are for negative m , while lines 28–42 are for positive m . $A_j^k / (-1)^j$ is calculated on line 12. We define a preprocessed function “`#define ODDEVEN(n) ((n & 1 == 1) ? -1 : 1)`”, which calculates $(-1)^n$ without using a power function. A_n^m is calculated on lines 19 and 33. $i^{|k-m|-|k|-|m|}$ is calculated on line 34 for the $m \geq 0$ case, and is always 1 for $m < 0$. Since $|k-m|-|k|-|m|$ is always an even number, it is possible to calculate $i^{|k-m|-|k|-|m|}$ as $-1^{(|k-m|-|k|-|m|)/2}$ and use the `ODDEVEN` function defined previously. Then, `anm`, `ajk`, and `sharedYnm` are multiplied to this result. The complex part $e^{im\beta}$ that was omitted in the first stage is calculated on lines 17–18 and 31–32 using the index $m-k$ instead of m ; `ere` is the real part and `eim` is the imaginary part. `CnmReal` and `CnmImag` in lines 21–22 and 36–37 are the real and imaginary parts of the product of all the terms described above. Finally, these values are multiplied to M_n^m in lines 23–26 and 38–41, where `sharedMnmSource[2*i+0]` is the real part and `sharedMnmSource[2*i+1]` is the imaginary part. We use the relation $M_n^{-m} = \overline{M_n^m}$ to reduce the storage of `sharedMnmSource`. Therefore, the imaginary part has opposite signs for the $m \geq 0$ case and $m < 0$ case. The real part of L_j^k is accumulated in `LnmTarget[0]`, while the imaginary part is accumulated in `LnmTarget[1]`.

Listing 4: Calculation of L_n^m in the M2L kernel

```

1  __device__ void m2l_kernel_core(float* LnmTarget,
2                                int j, float beta,
3                                float* sharedFactorial,
4                                float* sharedYnm,
5                                float* sharedMnmSource)
6  {
7      int i, k, m, n, jnkm;
8      float ere, eim, anm, ajk, cnm, CnmReal, CnmImag;
9      k = 0;
10     for(i = 0; i <= j; i++) k += i;
11     k = threadIdx.x - k;
12     // using pre-processed function ODDEVEN
13     ajk = ODDEVEN(j) * rsqrtf(sharedFactorial[j - k] * sharedFactorial[j + k]);
14     for(n = 0; n < numExpansions; n++){
15         for(m = -n; m < 0; m++){
16             i = n * (n + 1) / 2 - m;
17             jnkm = (j + n) * (j + n + 1) / 2 - m + k;
18             ere = cosf((m - k) * beta);
19             eim = sinf((m - k) * beta);
20             anm = rsqrtf(sharedFactorial[n - m] * sharedFactorial[n + m]);
21             cnm = anm * ajk * sharedYnm[jnkm];
22             CnmReal = cnm * ere;
23             CnmImag = cnm * eim;
24             LnmTarget[0] += sharedMnmSource[2 * i + 0] * CnmReal;
25             LnmTarget[0] += sharedMnmSource[2 * i + 1] * CnmImag;
26             LnmTarget[1] += sharedMnmSource[2 * i + 0] * CnmImag;
27             LnmTarget[1] -= sharedMnmSource[2 * i + 1] * CnmReal;
28         }
29         for(m = 0; m <= n; m++){
30             i = n * (n + 1) / 2 + m;
31             jnkm = (j + n) * (j + n + 1) / 2 + abs(m - k);
32             ere = cosf((m - k) * beta);
33             eim = sinf((m - k) * beta);
34             anm = rsqrtf(sharedFactorial[n - m] * sharedFactorial[n + m]);
35             cnm = ODDEVEN((abs(k - m) - k - m) / 2);
36             cnm *= anm * ajk * sharedYnm[jnkm];
37             CnmReal = cnm * ere;
38             CnmImag = cnm * eim;
39             LnmTarget[0] += sharedMnmSource[2 * i + 0] * CnmReal;
40             LnmTarget[0] -= sharedMnmSource[2 * i + 1] * CnmImag;
41             LnmTarget[1] += sharedMnmSource[2 * i + 0] * CnmImag;
42             LnmTarget[1] += sharedMnmSource[2 * i + 1] * CnmReal;
43         }
44     }
45 }

```

The functions in Listings 3 and 4 are called from an outer function shown in Listing 5. This function is similar to the one shown in Listing 2. The inputs are `deviceOffset` and `deviceMnmSource`. The output is `deviceLnmTarget`. The definitions are:

- `deviceOffset` contains the number of interacting cells, the offset of the particle index for each of these cells, and the 3D index of their relative positioning.
- `threadsPerBlockTypeB` and `maxM2LInteraction` are defined externally.
- `maxM2LInteraction` is the maximum size of the interaction list for the M2L, which is 189 for the present kernels.
- `offsetStride`, calculated on line 6, is the stride of the data in `deviceOffset`.

On line 8, the size of the cell is read from `deviceConstant[0]`, which resides in constant memory. On line 10, `LnmTarget` is initialized. Each thread handles a different coefficient in L_j^k . In order to do this, we must associate the `threadIdx.x` to a pair of j and k . `sharedJ` returns the index j when given the `threadIdx.x` as input. It is declared on line 11, initialized on lines 16–18, the values are calculated on lines 19–24, and then passed to `m2l_kernel_core()` on line 40. `sharedMnmSource` is the copy of `deviceMnmSource` in shared memory. It is declared on line 12 and the values are copied on lines 35–36 before it is passed to `m2l_kernel_core()` on line 41. `sharedYnm` contains the real spherical harmonics. It is declared on line 13 and its values are calculated in the function `m2l_calculate_ynm` on line 39 before they are passed to `m2l_kernel_core` on line 41. `sharedFactorial` contains the factorial for the given index and is declared on line 14 and its values are calculated on lines 25–29 before they are passed to `m2l_kernel_core` on line 41. On line 15, the number of interacting cells is read from `deviceOffset` and its value `numInteraction` is used for the loop on line 30. The offset of particles are read from `deviceOffset` on line 31, and the relative distance of the source and target cell are calculated on lines 32–34. On line 38, this distance is transformed into spherical coordinates using an externally defined function `cart2sph`. The two functions shown in Listings 3 and 4 are called on lines 39–41. Finally, the results in `LnmTarget` are copied to `deviceLnmTarget` on line 45.

Listings 1–5 are the core components of the present GPU implementation. We hope that the other parts of the open-source code that we provide along with this article are understandable to the reader without explanation.

6 Overview of Advanced Techniques

There are various techniques that can be used to enhance the performance of the treecode and FMM. The FMM presented in this article uses the standard translation operator for translating multipole/local expansions. As the order of expansion p increases, the calculation increases as $\mathcal{O}(p^4)$ for this method. There are alternatives that can bring the complexity down to $\mathcal{O}(p^3)$ [8] or even $\mathcal{O}(p^2)$ [17]. In the code that we have released along with this article, we have included an implementation of the $\mathcal{O}(p^3)$ translation kernel by [8] as an extension. We have omitted the explanations in this text, however, and consider the advanced reader able to self-learn the techniques from the literature to understand the code. Some other techniques that can improve the performance are the optimization

Listing 5: The entire M2L kernel

```

1  __global__ void m2l_kernel(int* deviceOffset, float* deviceLnmTarget,
2                               float* deviceMnmSource)
3  {
4      int i, j, k, ij, ib, numInteraction, jbase;
5      const int threadsPerBlock = threadsPerBlockTypeB;
6      const int offsetStride = 4*maxM2LInteraction+1;
7      float3 dist;
8      float boxSize = deviceConstant[0];
9      float rho, alpha, beta, fact;
10     float LnmTarget[2] = {0.0f, 0.0f};
11     __shared__ int sharedJ[threadsPerBlock];
12     __shared__ float sharedMnmSource[2 * threadsPerBlock];
13     __shared__ float sharedYnm[numCoefficients];
14     __shared__ float sharedFactorial[2 * numExpansions];
15     numInteraction = deviceOffset[blockIdx.x * offsetStride];
16     for(i = 0; i < threadsPerBlock; i++){
17         sharedJ[i] = 0;
18     }
19     for(j = 0; j < numExpansions; j++){
20         for(k = 0; k <= j; k++){
21             i = j * (j + 1) / 2 + k;
22             sharedJ[i] = j;
23         }
24     }
25     fact = 1.0;
26     for(i = 0; i < 2 * numExpansions; i++) {
27         sharedFactorial[i] = fact;
28         fact = fact * (i + 1);
29     }
30     for(ij = 0; ij < numInteraction; ij++){
31         jbase = deviceOffset[blockIdx.x * offsetStride + 4 * ij + 1];
32         dist.x = deviceOffset[blockIdx.x * offsetStride + 4 * ij + 2] * boxSize;
33         dist.y = deviceOffset[blockIdx.x * offsetStride + 4 * ij + 3] * boxSize;
34         dist.z = deviceOffset[blockIdx.x * offsetStride + 4 * ij + 4] * boxSize;
35         for(i=0;i<2;i++) sharedMnmSource[2 * threadIdx.x + i] =
36             deviceMnmSource[2 * (jbase + threadIdx.x) + i];
37         __syncthreads();
38         cart2sph(rho, alpha, beta, dist.x, dist.y, dist.z);
39         m2l_calculate_ynm(sharedYnm, rho, alpha, sharedFactorial);
40         m2l_kernel_core(LnmTarget, sharedJ[threadIdx.x], beta,
41             sharedFactorial, sharedYnm, sharedMnmSource);
42         __syncthreads();
43     }
44     ib = blockIdx.x * threadsPerBlock + threadIdx.x;
45     for(i=0;i<2;i++) deviceLnmTarget[2 * ib + i] = LnmTarget[i];
46 }

```

of the order of expansion for each interaction [10], the use of a more efficient M2L interaction stencil [18], and the use of a treecode/FMM hybrid, as suggested in [8]. It is needless to mention that the parallelization of the code for multi-GPU calculations [20, 22] is an important extension to the treecode/FMM on GPUs. Again, this is an advanced topic beyond the scope of this contribution.

When reporting the GPU/CPU speed up, it is bad form to compare the results against an unoptimized serial CPU implementation. Sadly, this is often done, which negatively affects the credibility of results in the field. For this contribution, we have used a reasonable serial code in C, but it is certainly not as fast as it could be. For example, it is possible to achieve over an order of magnitude performance increase on the CPU by doing single-precision calculations using SSE instructions with inline assembly code [31]. For those that are interested in the comparison between a highly tuned CPU code and highly tuned GPU code, we provide a highly tuned CPU implementation of the treecode/FMM in the code package that we release with this article.

7 Conclusions

This contribution is a follow-on from the previous *GPU Gems 3*, Chapter 31 [32], where the acceleration of the all-pairs computation on GPUs was presented for the case of the gravitational potential of N masses. We encourage the reader to consult that previous contribution, as it will complement the presentation we have given. As can be seen in the results presented here, the cross-over point where fast N -body algorithms become advantageous over direct, all-pairs calculations is in the order of 10^3 for the CPU and in the order of 10^4 for the GPU. Hence, utilizing the GPU architecture moves the cross-over point upwards by one order of magnitude, but this size of problem is much smaller than many applications require. If the application of interest involves, say, millions of interacting bodies, the advantage of fast algorithms is clear, in both CPU and GPU hardware. With our basic kernels, about $15\times$ speedup is obtained from the fast algorithm on the GPU for a million particles. For $N = 10^7$, the fast algorithms provide $150\times$ speedup over direct methods on the GPU. However, if the problem at hand requires small systems, smaller than 10^4 , say, one would be justified to settle for the all-pairs, direct calculation.

The main conclusion that we would like the reader to draw from this contribution is that constructing fast N -body algorithms on the GPU is far from a formidable task. Here, we have shown basic kernels that achieve substantial speedup over direct evaluation in less than 200 lines of CUDA code. Expert-level implementations will, of course, be much more involved, and would achieve more performance. But a basic implementation like the one shown here is definitely worthwhile.

We thank F. A. Cruz for various discussions that contributed to the quality of this article.

References

- [1] S. Aarseth. *Gravitational N-Body Simulations*. Cambridge University Press, 2003.
- [2] Andrew W. Appel. An efficient program for many-body simulation. *SIAM J. Sci. Stat. Comput.*, 6(1):85–103, 1985.

- [3] M. S. Bapat, L. Shen, and Y. J. Liu. Adaptive fast multipole boundary element method for three-dimensional half-space acoustic wave problems. *Engineering Analysis with Boundary Elements*, 33(8–9):1113–1123, August–September 2009.
- [4] J. Barnes and P. Hut. A hierarchical $O(N \log N)$ force-calculation algorithm. *Nature*, 324:446–449, December 1986.
- [5] J. E. Barnes. A modified tree code: Don’t laugh; it runs. *J. Comput. Phys.*, 87:161–170, 1990.
- [6] J. Board and K. Schulten. The fast multipole algorithm. *Computing in Science and Engineering*, 2(1):76–79, January/February 2000.
- [7] J. A. Board, Jr., J. W. Causey, J. F. Leathrum, Jr., A. Windemuth, and K. Schulten. Accelerated molecular dynamics simulation with the parallel fast multipole algorithm. *Chem. Phys. Lett.*, 198(1–2):89–94, 1992.
- [8] H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comput. Phys.*, 155:468–498, 1999.
- [9] E. Darve and P. Have. Efficient fast multipole method for low-frequency scattering. *J. Comput. Phys.*, 197:341–363, 2004.
- [10] H. Daschel. Corrected article: “An error-controlled fast multipole method”. *J. Chem. Phys.*, 132:119901, 2010.
- [11] K. C. Donepudi, J.-M. Jin, and W. C. Chew. A higher order multilevel fast multipole algorithm for scattering from mixed conducting/dielectric bodies. *IEEE Transactions on Antennas and Propagation*, 51(10):2814–2821, 2003.
- [12] F. Ethridge and L. Greengard. A new fast-multipole accelerated Poisson solver in two dimensions. *SIAM J. Sci. Comput.*, 23(3):741–760, 2001.
- [13] H. Fujiwara. The fast multipole method for integral equations of seismic scattering problems. *Geophys. J. Intl.*, 133:773–782, 1998.
- [14] Lena Gorelick, Meirav Galun, Eitan Sharon, Ronen Basri, and Achi Brandt. Shape representation and classification using the Poisson equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1991–2005, 2006.
- [15] L. Greengard and J.-Y. Lee. A direct adaptive Poisson solver of arbitrary order accuracy. *J. Comp. Phys.*, 125:415–424, 1996.
- [16] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.
- [17] N. A. Gumerov and R. Duraiswami. *Fast multipole methods for the Helmholtz equation in three dimensions*. Elsevier Series in Electromagnetism. Elsevier Ltd., 1st edition, 2004.
- [18] N. A. Gumerov and R. Duraiswami. Fast multipole methods on graphics processors. *J. Comp. Phys.*, 227(18):8290–8313, 2008.
- [19] Nail A. Gumerov and Ramani Duraiswami. A broadband fast multipole accelerated boundary element method for the three dimensional Helmholtz equation. *J. Acoust. Soc. Am.*, 125(1):191–205, 2009.
- [20] T. Hamada, T. Narumi, R. Yokota, K. Yasuoka, K. Nitadori, and M. Taiji. 42 TFlops hierarchical N-body simulations on GPUs with applications in both astrophysics and turbulence. In *SC ’09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–12, New York, NY, 2009. ACM.

- [21] A. Kawai, T. Fukushige, and J. Makino. \$7.0/Mflops astrophysical N -body simulation with treecode on GRAPE-5. In *Supercomputing '99: Proceedings of the 1999 ACM/IEEE conference on Supercomputing*, New York, NY, USA, 1999. ACM.
- [22] I. Lashuk, A. Chandramowlishwaran, H. Langston, T. Nguyen, R. Sampath, A. Shringarpure, R. Vuduc, L. Ying, D. Zorin, and G. Biros. A massively parallel adaptive fast-multipole method on heterogeneous architectures. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09*, pages 1–12, Portland, Oregon, November 2009.
- [23] Y. J. Liu and N. Nishimura. The fast multipole boundary element method for potential problems: A tutorial. *Engineering Analysis with Boundary Elements*, 30:371–381, 2006.
- [24] Yijun Liu. *Fast multipole boundary element method: Theory and applications in engineering*. Cambridge University Press, 2009.
- [25] J. Makino. Treecode with special-purpose processor. *Publ. Astron. Soc. Japan*, 43:621–638, 1991.
- [26] J. Makino. Yet another fast multipole method without multipoles–pseudoparticle multipole method. *J. Comput. Phys.*, 151:910–920, 1999.
- [27] J. Makino and M. Taiji. Astrophysical N -body simulations on GRAPE-4 special-purpose computer. In *Supercomputing '95: Proceedings of the 1995 ACM/IEEE conference on Supercomputing*, page 63, New York, NY, USA, 1995. ACM.
- [28] Junichiro Makino and Makoto Taiji. *Scientific Simulations with Special-Purpose Computers—the GRAPE Systems*. John Wiley & Sons Inc., 1998.
- [29] Herbert Nguyen, editor. *GPU Gems 3*. Addison-Wesley Professional, 2007. Available free online at <http://developer.nvidia.com/object/gpu-gems-3.html>.
- [30] N Nishimura. Fast multipole accelerated boundary integral equation methods. *Appl. Mech. Rev.*, 55(4):299–324, 2002.
- [31] K. Nitadori, K. Yoshikawa, and J. Makino. Personal communication.
- [32] Lars Nyland, Mark Harris, and Jan Prins. Fast N -body simulation with CUDA. In *GPU Gems 3*, chapter 31, pages 677–695. Addison-Wesley Professional, 2007.
- [33] V. Rokhlin. Rapid solution of integral equations of scattering theory in two dimensions. *J. Comp. Phys.*, 86(2):414–439, 1990.
- [34] C. Sagui and T. A. Darden. Molecular dynamics simulations of biomolecules: Long-range electrostatic effects. *Ann. Rev. Biophys. Biomol. Struct.*, 28:155–179, 1999.
- [35] Gary Taubes. The star machine. *Discover*, 18(6):76–83, 1997. available online at <http://discovermagazine.com/1997/jun/thestarmachine1148>.
- [36] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

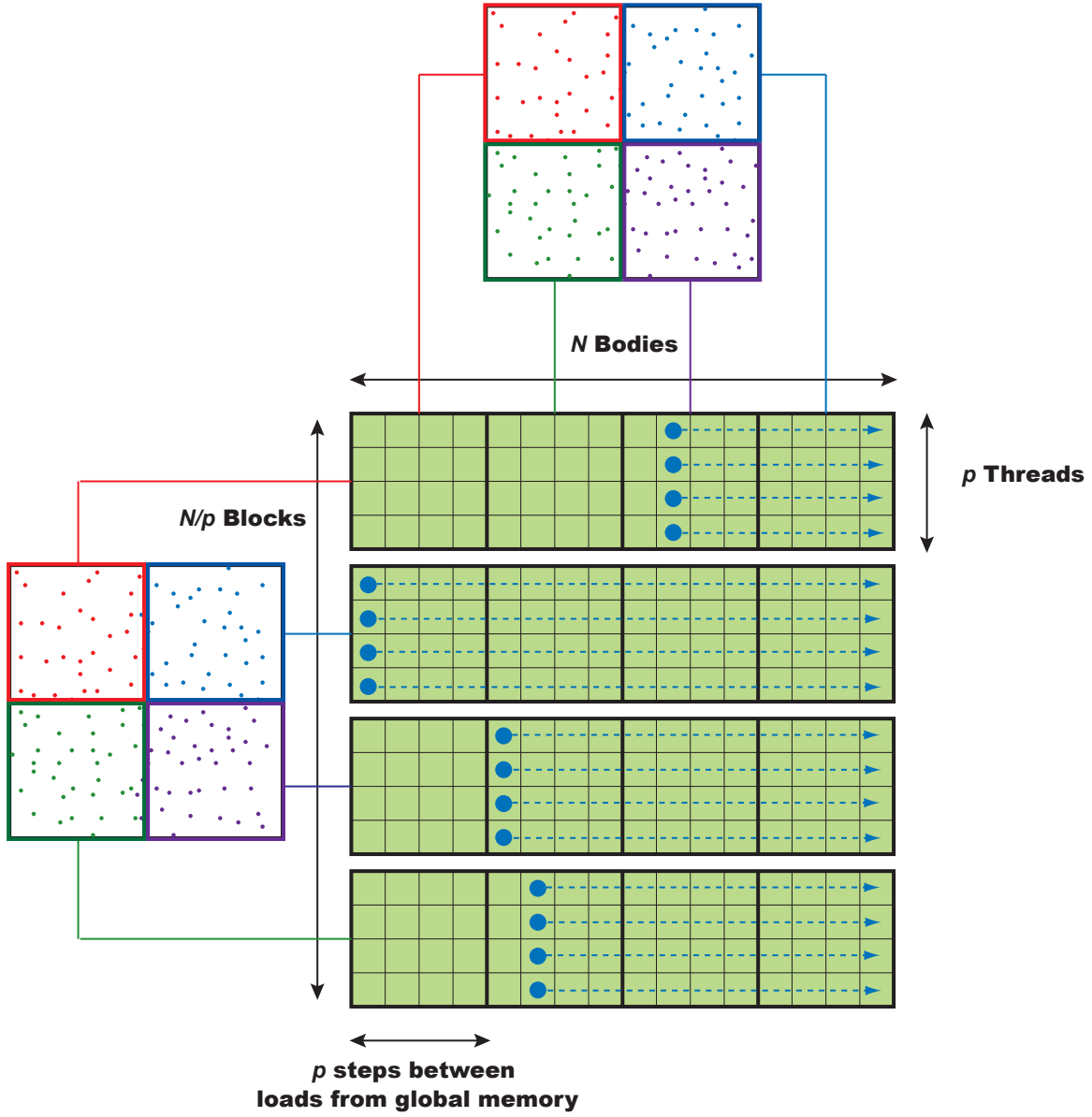


Figure 2: Thread block model of the direct evaluation on GPU; as in [32].

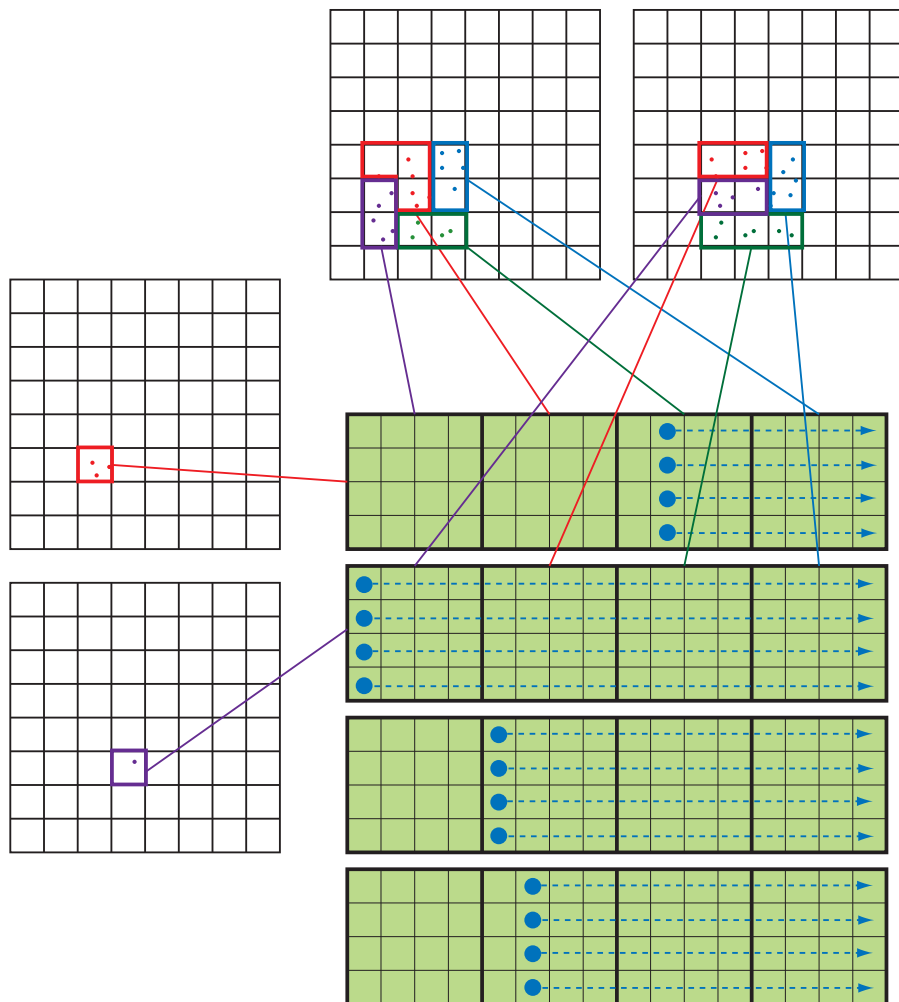


Figure 3: Thread block model of the particle-particle interaction on GPUs.

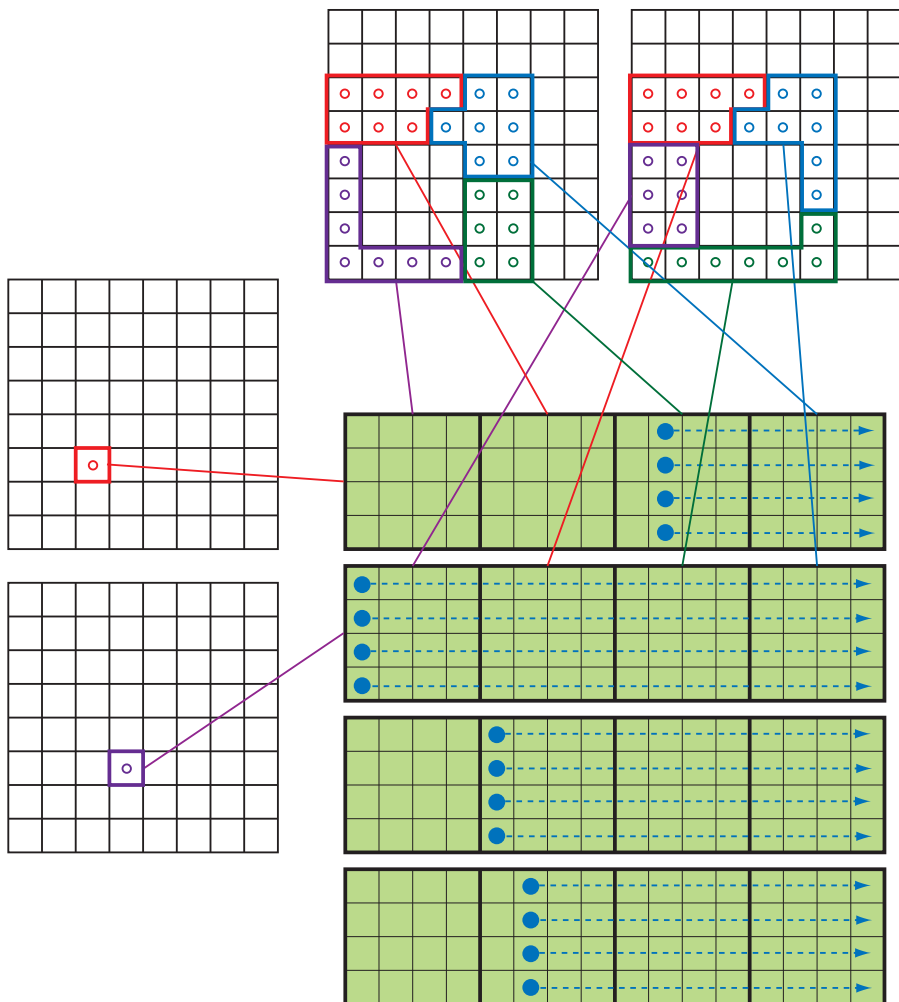


Figure 4: Thread block model of the cell-cell interaction on GPUs

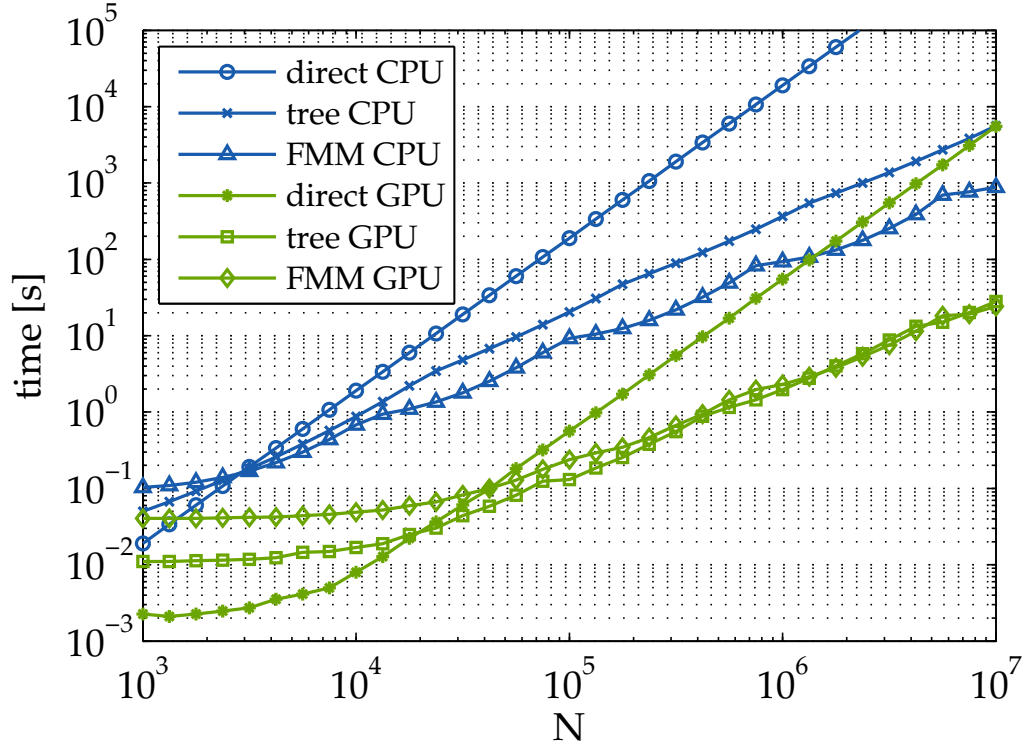


Figure 5: Calculation time for the direct method, treecode and FMM on CPU and GPU. (Normalized L^2 norm error of the force is 10^{-4} for both treecode and FMM).

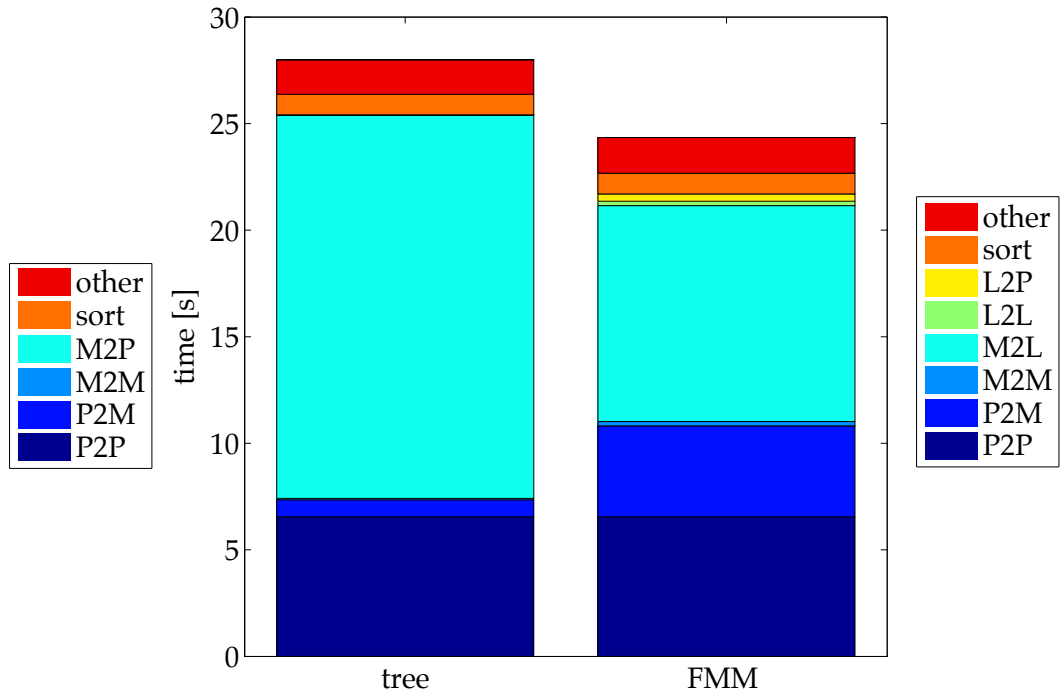


Figure 6: Breakdown of the calculation time for the treecode and FMM on GPUs using $N = 10^7$ particles.

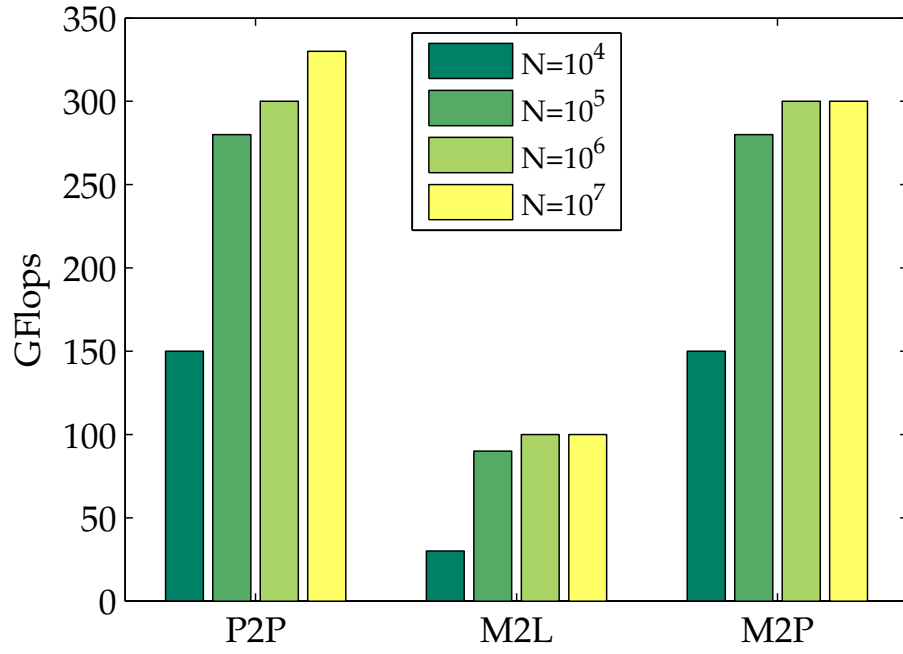


Figure 7: Actual performance in Gflop/s of three core kernels, for different values of N .